가변길이 윈도우와 빈도 가중치를 이용한 단어 의미 중의성 해소

Word Sense Disambiguation using Dynamic Sized Window and Frequency Weighting

저자
(Authors)

박상근, 최지연, 최기선
Sangkeun Park, Jeeyeon Choi, Key-Sun Choi

출처
(Source)

한국정보과학회 학술발표논문집 , 2014.12, 441-443 (3 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/Article/NODE06228673

APA Style

박상근, 최지연, 최기선 (2014). 가변길이 윈도우와 빈도 가중치를 이용한 단어 의미 중의성 해소. 한국정보과학회 학술발표논문집, 441-443.

가변길이 윈도우와 빈도 가중치를 이용한 단어 의미 중의성 해소

박상근, 최지연, 최기선O
한국과학기술원 지식서비스공학과, 한국과학기술원 전산학과O
sk.park@kaist.ac.kr, jeeyeon51@kaist.ac.kr, kschoi@kaist.eduO

Word Sense Disambiguation using Dynamic Sized Window and Frequency Weighting

Sangkeun Park, Jeeyeon Choi, Key-Sun ChoiO
KAIST Dept of Kowledge Service Engineering, KAIST Dept of Computer ScienceO

요 약

한국어에서 단어 의미 중의성 해소를 위해 다양한 연구가 진행되어 왔다. 본 논문에서는 가변길이 윈도우 사이즈 및 학습 데이터에서의 단어 의미 출현 빈도에 대한 가중치를 적용한 방법을 이용하여 단어 의미 중의성을 해소하는 방법을 제안하였다. SENSEVAL-2 한국어 데이터를 이용하여 학습하고 성능을 측정하였으며 10-fold 교차 검증 방법으로 실험한 결과, 최대 94.37%, 평균 86.39%의 우수한 성능을 보였다. 또한, 해당 결과를 같은 SENSEVAL-2 데이터를 이용한 이전 연구와 비교하여 해당 연구의 우수성을 입증하였다.

1. 서 론

자연언어처리는 검색엔진, 번역서비스 뿐만 아니라 소셜네트워크 상에서의 사용자의 감성 분석 등, 매우 다양한 분야에서 널리 쓰이고 있다. 또한 그 시대의 사회와 문화를 반영하는 언어적 특성으로 인해, 계속해서 변화하는 언어에 맞게 해당 연구도 계속해서 발전하고 있다. 본 논문에서는 자연언어처리에서 발생하는 여러 가지 문제 중, 의미 분석 단계에서 발생하는 단어 의미 중의성 해소(word sense disambiguation)를 다룬다. 단어 의미 중의성 해소는 하나의 단어가 여러 가지의 의미로 해석되는 문제를 해결하는 것을 의미한다. 본 논문에서는 교사 학습(supervised learning)으로, 가변길이 윈도우 사이즈와 빈도 가중치를 이용한 단어 의미 중의성 해소 방법을 제안하였다. 2장에서는 단어 의미 중의성 해소를 위한 국내외 관련 연구에 대하여 정리하고, 3장에서는 본 논문에서 제안하는 한국어 단어 의미 중의성 해소 방법에 대하여 설명한다. 4장에서는 본 논문에서 제안한 방법에 대한 성능을 평가 및 결과를 기술하며, 마지막으로 5장에서 결론 및 본 연구의 한계와 그 개선 방안을 제안한다.

2. 관련 연구

단어 의미 중의성 해소를 위한 방법으로, 크게 교사 학습 방법(Supervised learning), 자율 학습 방법(Unsupervised learning)이 있다. 교사 학습을 이용한 방법은 이미 단어의 의미를 알고 있는 단어들을 학습 데이터로 사용하여 단어 의미 중의성을 해소할 수 있다. 자율 학습 방법으로는 단어의 의미를 미리 알지 못하는 순수한 원시 말뭉치만을 이용하여, 동형이의어의 단어 의미 중의성을 해소한다[1].

교사 학습 방법을 통한 한국어 단어 의미 중의성 해소 방법으로, [2]에서는 고정길이의 윈도우를 사용하여 중의성을 해소하고자 하였고, 의미 분별 대상 단어를 기준으로 인접해있을수록 의미를 구분하는 데 중요한 정보를 담고 있을 확률이 높으며 윈도우 크기를 좌우 5어절로 고정했을 때에 중의성 해소에 결정적 영향을 미칠 수 있는 단어가 포함될 확률이 약 97.8%에 달함을 밝혔다. [3, 4]에서는 문장에 따라 가변적인 크기의 문맥을 사용하는 의미 분석 방법을 제안하였고, 고정 크기 문맥을 사용한 경우에 비해 향상된 결과를 보였다.

본 논문에서는 기존의 윈도우 크기를 이용한 교사 학습 방법을 개선하여, 더 빠르고 정확한 단어 의미 중의성 해소를 시도하였다. 또한 본 논문에서 이용한 SENSEVAL-2 데이터를 이용하여 단어 의미 중의성 해소를 시도한 관련 연구와 비교하여, 본 논문의 우수성을 입증하였다.

3. 단어 의미 중의성 해소

3.1 가변길이 윈도우

중의성을 가지는 단어의 의미를 결정짓는 문맥의 범위를 윈도우라고 한다. 윈도우의 크기에 따라 중의성 해소에 도움이 되기도 하고 방해가 되기도 한다. 윈도우가 너무 작을 경우 정보량이 부족하고 지나치게 많을 경우는 방해 요소가 되기도 한다. 따라서 적절한 크기의 윈도우를 사용하는 것은 중요한 일이다.

본 논문에서는 윈도우의 크기를 가변적으로 설정한다. 의미를 분별하고자 하는 대상이 체언일 때 좌우 최초로 나오는 용언까지를 윈도우로 결정한다. 이는 함께 쓰이는 용언이 해당 체언의 의미를 분별하는데 유용한 정보로 사용될 수 있기 때문이다. 또한 윈도우의 크기가 지나치게 커질 경우 방해 요소들이 추가될 가능성이 있으므로 최대 좌우 5어절까지로 제한한다. 이는 [2]에서 밝힌, 우리말의 단어 의미 중의성 해소에 일반적으로 좌우 5어절을 사용한다는 내용을 기반으로 한다.

예를 들어, “덜 익은 감을 먹고 배탈이 났다.”라는 문장에서 ‘감’의 의미를 분별하고자 할 때 윈도우를 결정하는 과정은 다음과 같다. 먼저 왼쪽으로 탐색할 때

최초로 나오는 용언은 '익'(익다)이다. 다음으로 오른쪽을 탐색해보면 최초로 나오는 용언이 '먹'(먹다)이므로 윈도우는 '익은 감을 먹'이 된다.

3.2 의미 분별 모델

가변길이 윈도우를 적용하여 다음과 같이 의미 분별 대상 단어의 의미를 결정한다.

$$\mathrm{argmax}_{s_j \in S}\left(\sum_{w_i \in W} \delta_{ij}\right) \qquad \delta_{ij} = \begin{cases} 1, w_i \in W(s_j) \\ 0, w_i \notin W(s_j) \end{cases}$$

s_j는 의미가 부착된 단어이고, w_i는 의미를 분별하고자 하는 대상 단어의 윈도우에 속한 단어들이다. δ_{ij}는 w_i가 s_j의 윈도우에 포함된 단어들의 집합에 포함될 경우 1, 그렇지 않을 경우 0을 반환하는 함수이다. 따라서 이 모델은 의미가 부착된 단어들마다 각 단어의 윈도우에 속한 단어들의 집합을 기준으로 하여, 의미를 분별하고자 하는 단어의 윈도우에 속한 단어들이 기준과 비교했을 때, 가장 다수 일치하는 단어의 의미를 대상 단어의 의미로 채택하는 방식이다. '말'이라는 단어의 뜻이 동물을 뜻하는 '말1'과 사람이 생각이나 느낌을 표현하기 위해 사용하는 '말2'로 나뉘는 경우를 예로 들면 s_1은 '말1'이 되고, s_2는 '말2'가 된다. 주어진 말뭉치에서 s_1, s_2 각각에 대해 윈도우에 속하는 단어들을 추출하고 그 단어들의 집합을 각각 $W(s_1)$, $W(s_2)$라고 한다. 그리고 의미를 분별하고자 하는 '말'이 포함된 문장에서 해당 '말'의 윈도우에 속한 단어들을 추출하여 $W(s_1)$과 $W(s_2)$의 단어집합과 비교한다. 만약 $W(s_1)$의 단어들 중 일치하는 수가 더 많은 경우, 주어진 '말'은 동물을 뜻하는 말이라고 판단한다.

상기 모델을 이용하였을 때, s_j에 대한 결과식 $\sum_{w_i \in W} \delta_{ij}$와 s_k에 대한 결과식 $\sum_{w_i \in W} \delta_{ik}$의 값이 같게 나오는 경우에 대해 argmax를 정할 수 없는 한계점이 있으므로, 이러한 경우를 보완하기 위하여 단어(s_j) 자체의 빈도를 고려하도록 아래와 같이 개선하였다.

$$\mathrm{argmax}_{s_j \in S}\left(\sum_{w_i \in W} \delta_{ij} + P(s_j)\right) \qquad P(s_j) = \frac{c(s_j)}{\sum_{s_i \in S} c(s_i)}$$

$P(s_j)$는 s_j의 빈도로, 주어진 말뭉치에서 s의 형태를 가진 단어들 중 s_j의 의미를 가진 단어의 비율이다. 0에서 1사이의 값인 $P(s_j)$와 0 이상의 정수인 $\sum_{w_i \in W} \delta_{ij}$가 합으로 연결되어 있기 때문에 $\sum_{w_i \in W} \delta_{ij}$와 $\sum_{w_i \in W} \delta_{ik}$가 같은 값을 가질 경우에만 $P(s_j)$(혹은 $P(s_k)$)가 의미를 가진다. 즉, 수정된 모델은 $\sum_{w_i \in W} \delta_{ij}$와 $\sum_{w_i \in W} \delta_{ik}$가 같은 값이 나오는 경우에 대해서만 이전 모델과 차이를 보이며, 최종 결과로는 s_j와 s_k 중 말뭉치에 더 많이 등장한 의미를 채택한다.

4. 실험 및 평가

4.1 실험 환경

본 논문에서는 SENSEVAL-2의 한국어 학습 데이터를 이용하여 제안된 단어 의미 중의성해소의 정확도를 측정하였다. SENSEVAL-2의 한국어 데이터는 '밤(141개)', '바람(141개)', '의사(230개)', '거리(182개)', '자리(130개)', '점(120개)', '말(120개)', '목(131개)', '눈(196개)', '손(132개)' 등 총 10개(1523개)의 동형이의어로 구성되어 있다. 본 논문에서의 단어 의미 중의성 해소의 정확도는 전체 어휘 중에서 정확하게 의미를 구분한 단어의 개수의 백분율 값으로 계산하였다.

$$\text{Accuracy}(\%) = \frac{n(\mathrm{argmax}_{s_j \in S}(\sum_{w_i \in W} \delta_{ij} + P(s_j)) = s)}{n(w)} \times 100$$

4.2 실험 방법

본 논문에서 제안한 단어 의미 중의성 해소 알고리즘의 성능을 평가하기 위하여, SENSEVAL-2를 이용하여 통계적으로 가장 많이 나타난 의미를 해당 단어의 의미로 결정하는 방법(Most Frequent Class) 및 어휘의미망의 의미 관계를 이용하여 어의 중의성을 해소한 방법[1]과 비교하여 성능을 측정한다.

SENSEVAL-2의 트레이닝셋과 테스트셋의 양이 매우 적으므로, 성능측정의 통계적 신뢰도를 높이기 위하여 트레이닝셋과 정답을 알고 있는 테스트셋을 모두 합친 뒤 10-fold 교차 검증 방식으로 성능을 테스트하였다.

4.3 실험 결과

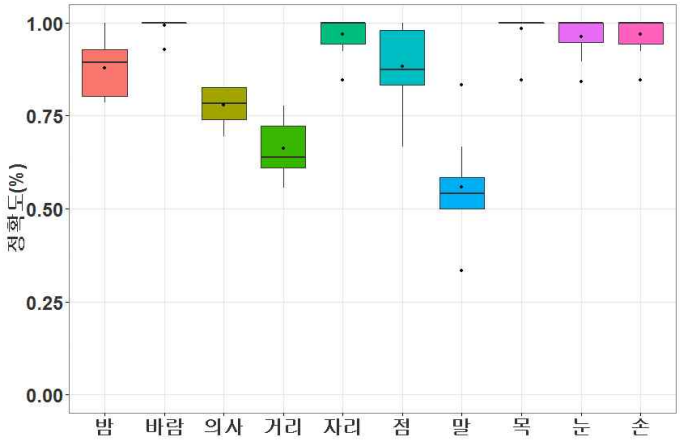

그림 1 각 단어별 10-fold 교차 검증 결과

본 논문에서 제안한 가변길이 윈도우 사이즈와 빈도 가중치를 이용한 알고리즘으로 SENSEVAL-2 한국어 데이터의 각 단어별 10-fold 교차 검증을 실행한 결과는 그림 1과 같다.

10-fold 교차 검증 결과에서 최대 정확도만 고려하였을 때, '의사', '거리', '말'을 제외한 7개 중의성 단어에서 100%의 정확도를 보였으며, 평균 94.4%로 매우 높은 정확도를 보이고 있다. 또한 각 중의성 단어의

중의성 어휘	정확도(%)						
	어휘의미망의 의미관계를 이용한 어의 중의성 해소 [5]					가변길이 윈도우 및 빈도 가중치 적용	
	MFC	기본 알고리즘	개선방안 1	개선방안 1+2	개선방안 1+2+3	평균	최대
밤	71.29	77.23	77.23	77.23	77.23	87.86	100
바람	98.98	98.98	96.94	96.94	94.90	99.29	100
의사	62.42	56.36	87.27	89.70	88.48	77.83	82.61
거리	53.44	47.33	47.33	68.70	74.05	66.11	77.78
자리	89.11	95.05	96.04	96.04	96.04	96.92	100
점	89.90	90.91	88.89	89.90	94.95	88.33	100
말	34.65	46.53	54.46	54.46	65.35	55.83	83.33
목	99.00	97.00	94.00	94.00	96.00	98.46	100
눈	93.98	93.98	93.98	94.74	94.74	96.32	100
손	97.73	93.18	97.73	97.73	98.48	96.92	100
평균	78.29	78.21	83.29	86.22	88.11	86.39	94.37

표 1 MFC 및 어휘의미망의 의미관계를 이용한 어의 중의성 해소 방법[5]과의 정확도 비교

10-fold 교차 검증 결과의 평균만 고려하였을 때, 평균 86.39%로 이 또한 상당히 우수한 정확도를 보인다.

해당 결과를 SENSEVAL-2에서 통계적으로 가장 많이 나타난 의미를 해당 단어의 의미로 결정하는 방법(Most Frequent Class, MFC) 및 어휘의미망의 의미 관계를 이용하여 어의 중의성을 해소한 방법[1]과 비교한 결과는 표1과 같다.

각 중의성 단어의 평균 정확도만을 고려하였을 때, 전체 평균은 MFC에 비하여 비약적인 정확도 향상을 보이지만, 이전 연구 [5]의 개선 알고리즘에 비해서는 약 2% 가량의 정확도가 떨어지는 결과를 보였다. 하지만 최대 정확도는 그에 비해 훨씬 우수한 성능을 보이고 있으며, 상대적으로 현저히 낮은 정확도를 보이는 '의사', '거리', '말' 에서도 기존의 연구보다 훨씬 나은 정확도를 보인다.

5. 결론 및 향후 연구

본 논문에서는 가변길이 윈도우 사이즈 및 학습 데이터에서의 단어 빈도 가중치를 이용한 단어 의미 중의성 해소 방법을 제안하였다. SENSEVAL-2 한국어 데이터셋의 크기가 매우 작은 단점을 극복하기 위하여, 10-fold 교차 검증을 이용하여 해당 알고리즘의 정확도를 측정하였으며, SENSEVAL-2 를 이용하여 단어 의미 중의성 해소 방법을 평가한 이전 연구의 결과와 비교하여, 본 논문에서 제안한 방법이 상대적으로 우수함을 보였다.

본 논문에서 제안한 방법은 따로 중의성을 해소하고자 하는 단어가 포함된 문장 뿐 아니라, 윈도우 크기가 허용하는 범위 내의 모든 단어를 고려하였다. 본래 해당 단어가 포함된 문장만을 대상으로 가변 길이 윈도우 사이즈를 적용하려 하였으나, 본 연구에서 사용한 파이썬 기반 자연어처리 라이브러리인 KoNLPy의 문장 단위 구분의 성능이 매끄럽지 못하여 해당 문장을 벗어난 단어까지 고려하게 되었다. 단어 의미 중의성의 단어가 포함된 문장만을 대상으로 본 연구를 적용한다면 좀 더 빠른 연산 속도와 보다 높은 정확도를 보장할 수 있다.

또한, SENSEVAL-2 한국어 데이터셋에서 동형이의어들의 태그된 의미가 균등하게 분포되어 있지 않고, 한쪽으로 치우친 경우가 많아 기본적으로 MFC의 결과가 높은 상태이다. 향후 연구로, 다른 데이터셋을 이용하여 본 논문에서 제시한 단어 의미 중의성 해소 방법을 평가하는 것이 추가로 필요하다.

참고 문헌

[1] Schütze, HinrichH., "Automatic word sense discrimination.", Computational linguistics 24.1, p97-123. 1998.

[2] 허정, 옥철영, "사전의 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템", 한국정보과학회, 정보과학회 논문지, 소프트웨어 및 응용, 제28권 제9호 p.68-698, 2001.

[3] 이현아, "가변 크기 문맥과 거리가중치를 이용한 동형이의어 중의성 해소", 한국마린엔지니어링학회, 제38권 제4호 p.444~450, 2014.

[4] 박기태, 이태훈, 황소현, 이현아, "가변길이 윈도우를 이용한 통계 기반 동형이의어의 중의성 해소", 제24회 한글 및 한국어 정보처리 학술대회, pp.40-46, 2012.

[5] 김민호, 권혁철. "한국어 어휘의미망의 의미 관계를 이용한 어의 중의성 해소." 정보과학회논문지: 소프트웨어 및 응용, 제38권 제10호 p554-564. 2011.